UNITED STATES ENVIRONMENTAL PROTECTION AGENCY
NATIONAL CENTER FOR ENVIRONMENTAL ASSESSMENT
RESEARCH TRIANGLE PARK, NC 27711

OFFICE OF
RESEARCH AND DEVELOPMENT

DATE: October 19, 2001

TO: Annie Jarabek, NCEA
Kevin Crofton, NTD, NHEERL

FROM: Allan H. Marcus, NCEA-RTP Allan H Marcus

RE: Explanation of Re-analyses of Perchlorate Hormone Data from the 1998 External Review Draft

I've reviewed the ANOVA runs in Appendices 1 - 8 in sufficient detail, given the short time, to believe that they are correct and represent the approaches we previously discussed. The results of these comparisons may be used to evaluate the sensitivity of the analyses to different approaches for collapsing sums of squares in ANOVA models. In general, the "conservative" and "liberal" approaches give identical results on the statistical significance of the dose effects from which NOAEL and LOAEL were derived for these data sets. It might be helpful to summarize the results in a manner that emphasizes the similarity of findings.

The "conservative" approach uses a standard ANOVA model for factorial designs. The main effects and interaction mean squares are compared to an error mean square which is calculated from the 'full model' residuals including all main effects and interactions, whether or not they are significant. Since the null hypotheses are specified *a priori*, the nominal P-values and confidence levels should accurately characterize the uncertainty of the estimated dose effects, and by implication should validate the groupings in the Duncan procedure from which NOAEL and LOAEL may be derived.

The "liberal" approach also starts with a standard ANOVA model for factorial designs. The main effects and interaction mean squares are compared to an error mean square which is calculated from the 'full model' residuals including all main effects and interactions, whether or not they are significant. If non-significant interaction terms are found, a smaller model with these terms omitted is fitted to the data, retaining only main effects and interactions previously found significant. The contribution of the non-significant effects removed from the second-stage model are added to the residual sum of squares in the denominator, as are the corresponding degrees of freedom of the effects removed to the residual d.f., as shown below.

Since the null hypotheses in the second stage are not specified *a priori*, the nominal P-values and confidence levels may not accurately characterize the uncertainty of the estimated dose effects, as with most other stepwise modeling procedures in which preliminary testing is carried out. The groupings in the Duncan procedure from which NOAEL and LOAEL are derived may have P-values different from the nominal levels. The choice to exclude effects with first-stage P-values > 0.05 is an acceptable if harsh choice. Many common stepwise regression programs use (nominal) P-to-enter and P-to-remove values of 0.15, a less stringent criterion. There are a few cases in the results in Draft Version 2 in which a P-to-remove of 0.15 might have led to different second-stage results, although I suspect this would have had little effect on the NOAEL and LOAEL.

Reversing the order of entry of the terms had little effect on Type I sums of squares in the model. There was little difference between Type I and Type III sums of squares, and no change in the significance of the findings. The conclusions appear reasonably insensitive to the model.

Dr. Haseman has criticized the EPA analyses in which important (presumably significant) sources of variation were pooled with the full-model residual mean square error. The example he cites (TSH at 120 days, p. 41 of Crofton, 1998b) goes from the full model (in SAS notation)

$$Y = \text{Gender} \mid \text{Treatment} = \text{Gender} + \text{Treatment} + \text{Gender*Treatment}$$

to a reduced model

$$Y = \text{Treatment}.$$

The denominator MS in the full model is based on residual error, MSE = SSE / dfe. The denominator MS in the reduced model (using u$v to mean the effect of u after adjusting for v) is

$$\text{denom.MS} = (\text{SSE} + \text{SS}(\text{Gender \$ Mean}) + \text{SS}(\text{Gender*Treatment \$ Gender, Mean})) / (\text{dfe} + \text{df}(\text{Gender}) + \text{df}(\text{Gender*Treatment})).$$

We agree that the use of the reduced model is not appropriate when, as in the example shown, the highly significant Gender effect is pooled in with the other terms. The Gender main effect is an important source of variation and assessment of the Treatment effect should be adjusted for Gender. It is less obvious that the non-significant interaction term should be retained.

The model acceptable to Dr. Haseman is the full model, whose ANOVA table is:

Effect	df	SS	MS	F	P
Gender	1	883.7	883.7	310.1	<0.0001
Treatment	3	65.2	21.7	7.63	0.0002
Gender * Treatment	3	12.0	4.0	1.40	0.2497
Error	71	202.3	2.85		

The reduced model from (p. 79, Crofton, 1998b) that Dr. Haseman finds unacceptable produces:

Effect	df	SS	MS	F	P
Treatment	3	65.7	21.9	1.50	0.2224 NS
Error	75	1097.4	14.63		

An alternative approach, by way of sensitivity analysis, is Y = Gender Treatment, which retains the highly significant factor Gender in the full model as a source of variability:

Effect	df	SS	MS	F	P
Gender	1	883.7 ?	883.7	304.7	<0.0001
Treatment	3	65.2 ?	21.7	7.48	0.000?
Error	74	214.3 ?	2.90		

? Means that the actual values were not available, but are likely to be similar to the full model.

This eliminates the non-significant Gender * Treatment effect and pools its effect with the denominator sum of squares = 202.3 + 12.0 = 214.3. This approach, which we call the "liberal" approach, pools only the non-significant interactions and main effects (depending on the value of the P-to-remove) with the error mean square. In this example, it makes little difference whether the P-to-remove for non-significant effects is 0.05, as in the following examples, or a larger value such as 0.15 commonly used in stepwise regression models. The difference is negligible because the effects retained are strong, with only slight reductions in the F values. In cases shown below, the effects in the full model are more marginal and different "liberal" models could be fitted with different criteria for P-to-remove. In any event, this approach would avoid the removal of *significant* sources of variation from the fitted model, a matter of concern to Dr. Haseman.

Sample tables:

KEY to nominal statistical significance:

****	$P \leq 0.0001$	
***	$0.0001 < P \leq 0.0010$	
**	$0.0010 < P \leq 0.01$	
*	$0.01 < P \leq 0.05$	
+	$0.05 < P \leq 0.10$	'marginal' significance
-	$0.10 < P \leq 0.15$	'removable effect'
NS	$P > 0.15$	not significant

APPENDIX 1:
Thyroid Lumen Area: Conservative analysis

Effect	d.f.	P (significance)
Block	1	0.0015 **
Dose	4	0.0016 **
Block*Dose	4	0.3339 NS
Gender	1	0.9160 NS
Block*Gender	1	0.0557 +
Dose*Gender	4	0.8105 NS
Block*Dose*Gender	4	0.0694 +
error	80	

Thyroid Lumen Area: Liberal analysis with P-to-remove of 0.05

Effect	d.f.	P (significance)
Block	1	0.0019 **
Dose	4	0.0008 ***
error with removed effects	94	

Duncan Groups of Treatment Effect in Conservative vs. Liberal Tests

```
Dose                            0      1      0.3           3      10
Conservative Duncan grouping    AAAAA         CCCCCCCCCCCCCCC
                                       BBBBBBBBBBBBBBB

Liberal Duncan grouping         AAAAA         CCCCCCCCCCCCCCC
                                       BBBBBBBBBBBBBBB


Mean                            318    273    260           222    211
```

An alternative model could be developed by retaining the marginally significant interactions using
a P-to-remove of 0.15. This is a less liberal approach and could affect the decision as to whether or not the marginally significant interaction terms should be retained. Some analysts would insist on retaining the Gender main effect because interactions with Gender may be significant, even though Gender is not a significant main effect in the full model. As with any stepwise procedure, the statistician must make some subjective judgements on model strategy and tactics. The

ANOVA model for this alternative is

Y = Block Dose [Gender?] Block*Gender Block*Dose*Gender

APPENDIX 4: Rat subchronic study with T3

Day 15: No reason to remove any effects
Day 90: Gender main effect NS, Gender*Dose P = 0.0038, Dose P < 0.0001, so that analyses for Dose stratified by Gender were reported, a reasonable procedure.

Day 120: Conservative approach with T3 (page 103)

Effect	d.f.	P (significance)
Dose	3	0.0005 ***
Gender	1	0.0.0374 *
Dose*Gender	3	0.4281 NS
error	71	

Day 120: Liberal approach with T3 (page 104)

Effect	d.f.	P (significance)
Dose	3	0.0005 ***
Gender	1	0.0.0371 *
error	74	

Duncan Groups of Treatment Effect in Conservative vs. Liberal Tests

Dose	0.05	CONTROL	1.00	10
Conservative Duncan grouping	AAAAAAAAAAAAAAAA			B
Liberal Duncan grouping	AAAAAAAAAAAAAAAA			B
Mean	215	214	213	191

etc. etc. etc.

Close Calls

There are only a few places in the appendices where the P-values are close to the boundary between significance and non-significance for P-to-remove values between 0.05 and 0.15, and the groupings may be sensitive to an alternative liberal model. They are:

Appendix 1, PND5, Block*Gender and Block*Gender*Dose are marginally significant
In the Conservative Model and would be retained with P-to-remove = 0.15
For PND90, Dose and Dose*Gender are retained by P-to-remove of 0.15, but not 0.05.

Appendix 4, Day 120, if Gender before Dose, then P for gender is 0.0411 (Type I), which close to the nominal P-to-remove of 0.05.

Appendix 5, Day 90, Gender and Gender*Dose would be retained in the Liberal approach when P-to-remove = 0.15.

Appendix 6, Day 90, Gender*Dose is marginally significant and would be retained in the Liberal approach when P-to-remove = 0.15.

In general, the effects are either strongly significant or strongly insignificant, so the Conservative and Liberal approaches give almost the same results because the Liberal approach pools only small factor sums of squares with the full model pure error sum of squares.

These decisions require a close collaboration among the toxicologist, the risk assessor, and the statistician in order to determine whether statistically problematic factors and interaction terms have biological significance and should therefore be retained or dropped in reporting the results of the analyses. B